Raja Gond

Microsoft Research India | IIT Bombay'23
rajagond.github.io raja.gond@outlook.com github.com/rajagond Google Scholar

EDUCATION

Indian Institute of Technology (IIT) Bombay, Mumbai, MH, India

- Bachelor of Technology (B.Tech) in Computer Science and Engineering — Jul 2019 – May 2023
  - Top CS Program in India. With **Honors**. Cumulative GPA: **8.6 / 10.0**
  - [redacted]
  - [redacted]

PUBLICATIONS

PREPRINTS

[3] **TokenWeave**: Efficient Compute–Communication Overlap for Distributed LLM Inference
Raja Gond, Nipun Kwatra, and Ramachandran Ramjee
DOI: 10.48550/arXiv.2505.11329 (**4 citations**, Under Submission at **MLSys 2026**)
Code: `https://github.com/microsoft/tokenweave` (66 stars)
Presented as a poster at the *Microsoft Research India Academic Summit*, June 24, 2025.

[2] **LLM-**[redacted]
Raja Gond, [redacted], Ramachandran Ramjee, and Ashish Panwar
DOI: to be released (Under Submission at **MLSys 2026**)
Code: `https://github.com/microsoft/llm-`[redacted] (to be released)

[1] **emucxl**: an emulation framework for CXL-based disaggregated memory applications
Raja Gond and Purushottam Kulkarni
DOI: 10.48550/arXiv.2404.08311 (1 Citation, Apr 2024)
Code: `https://github.com/cloudarxiv/emucxl` (17 stars)

RESEARCH EXPERIENCE

Microsoft Research Lab, Bengaluru, KA, India

- Pre-doctoral Researcher (Research Fellow), AI-Infrastructure — Jul 2023 – Present
  - Project: **Compute–Communication Overlap for Efficient Distributed LLM Inference**
  - Advisors: Dr. Nipun Kwatra and Dr. Ramachandran Ramjee
  - **TokenWeave:**
    - Proposed TokenWeave, a system for efficient compute–communication overlap in Tensor Parallel LLM inference, achieving up to $\mathbf{1.28}\times$ latency and up to $\mathbf{1.26}\times$ throughput improvements on $8\times$H100 GPUs over the vLLM baseline, outperforming solutions such as TileLink and NanoFlow on modern NVIDIA GPUs
    - Identified **RMSNorm (+Residual)**, a previously overlooked operation, as a non-trivial source of latency (6–8%)
    - Developed a novel fused **AllReduce–RMSNorm** kernel using Multimem-PTX instructions available on modern GPUs such as **Hopper and Blackwell** that performs as fast as AllReduce alone, makes RMSNorm overhead negligible, and uses only 2–6% of Streaming Multiprocessors (SMs), leaving most SMs available for compute
    - Designed a GPU-wave-aware, at most two-way token split compatible with mixed prefill and decode batches to overlap compute and communication (plus RMSNorm through the fused kernel) even for small token batches
    - A pull request is pending to integrate the fused kernel into SGLang (PR #10544). Some of the findings were later confirmed by an NVIDIA technical blog (enabling-fast-inference-and-resilient-training-with-nccl)
  - **Before TokenWeave: Overlap for Efficient Inference in Mixture-of-Experts (MoE) Models:**
    - Implemented **expert parallelism in vLLM** and showed its benefits over tensor parallelism for MoE architectures
    - Designed a lightweight signaling mechanism to initiate Direct Memory Access (DMA)-based partial GPU–GPU communication, freeing all SMs for computation and enabling effective compute–communication overlap
    - Achieved up to **20%** lower Mixtral-22B MoE FFN plus communication latency on $8\times$H100 in microbenchmarks
  - Project: **LLM-**[redacted]
  - Advisors: Dr. Ashish Panwar and Dr. Ramachandran Ramjee
    - [redacted]
    - [redacted]
    - [redacted]
  - Project: **Accelerating Multimodal Inference**
  - Advisors: Dr. Ashish Panwar, Dr. Nipun Kwatra, and Dr. Ramachandran Ramjee
    - Currently working on improving inference throughput for large-scale vision models through optimized GPU scheduling and concurrent execution of compute-bound image encoding and memory-bound decode tasks

UNDERGRAD RESEARCH EXPERIENCE

Dept. of Computer Science and Engineering, IIT Bombay, Mumbai, MH, India

- Undergraduate Researcher, SynerG Lab — Aug 2022 – Jun 2023
  - Project: **emucxl**: Emulation Framework and Access Library for CXL-Based Disaggregated Memory Systems
  - Advisor: Prof. Purushottam Kulkarni
    - ▹ Developed a user-space library coupled with a **NUMA-based CXL emulation backend** to enable rapid prototyping of disaggregated memory solutions via standardized CXL memory access
    - ▹ Conducted a literature survey on CXL standards and showed emucxl capabilities through practical use cases
  - Project: Persistent Memory (PMem) Applications [PDF, code]
  - Advisors: Prof. Purushottam Kulkarni and Prof. Umesh Bellur
    - ▹ Designed and implemented a robust reader-writer program on non-volatile memory that provides fault tolerance and efficient data access, using advanced array and pointer techniques
    - ▹ Explored **Persistent Memory Development Kit** libraries to understand PMem capabilities and analyzed performance differences between traditional and PMem-based Redis using real-world benchmarks

INDUSTRY EXPERIENCE

Morgan Stanley, Mumbai, MH, India

- Technology Analyst Intern, Investment Management Division — May 2022 – Jul 2022
  - ▹ Designed and implemented a Java utility library for translating MT Swift payment messages generated by a trading platform into enriched MX messages, streamlining the migration process to new messaging standards
  - ▹ Integrated MX format verification and conducted analysis of MT-MX equivalence and translation rules
  - ▹ Received an offer for a **full-time position** upon graduation, based on exemplary internship performance

TEACHING

Undergraduate Teaching Assistant, Dept. of Computer Science and Engineering, IIT Bombay

- Computer Networks + Lab (CS224/CS252) — Spring'23
  - ▹ Instructor: Prof. Bhaskaran Raman
  - ▹ Evaluated exam answer sheets, explained concepts, and resolved doubts for over **200 CSE sophomores**
- Operating Systems + Lab (CS347/CS333) — Fall'22
  - ▹ Instructors: Prof. Purushottam Kulkarni and Prof. Umesh Bellur
  - ▹ Designed lab assignments, addressed students' doubts during lab sessions and online, proctored theory and lab exams, and evaluated answer scripts and lab coding assignments for a batch of over **180 CSE juniors**
- Computer Systems (Bootcamp) — Summer'22
  - ▹ Instructors: Prof. Purushottam Kulkarni and Prof. Mythili Vutukuru
  - ▹ Involved in the design of weekly assignments and in asynchronous doubt-solving to aid self-paced learning

MENTORSHIP

Department Academic Mentor, Student Mentorship Program, IIT Bombay — Jul 2022 – Apr 2023

- ▹ Selected out of **70+** applicants through a rigorous procedure based on SoP, interviews, and peer reviews
- ▹ Mentored students with academic or general concerns to help ease their transition into the CSE department

SERVICE

Artifact Evaluation Committee: OSDI/ATC'25, SOSP'25, EuroSys'26

- ▹ Evaluated two research artifacts for each conference, wrote reviews, and participated in committee discussions

AWARDS & SCHOLARSHIPS

- Microsoft Global Hackathon 2023: Executive Challenge First Prize Award — Sep 2023
  Hack for the Microsoft Cloud in the Era of AI (Idea: Microsoft Confidential)
  Collaborated closely with the Hackathon teammates spread across global Microsoft offices to develop an innovative solution that enhances cloud infrastructure capabilities and presented it to the Microsoft Cloud + AI leadership
- [redacted]
- [redacted]

SELECTED ACADEMIC PROJECTS

Dept. of Computer Science and Engineering, IIT Bombay

- SCLP: Compiler for C-like Language — Spring'22
  Guide: Prof. Uday Khedker — Implementation of Programming Languages
  - ▹ Built a compiler to generate Abstract Syntax Tree, Three-Address Code, and corresponding assembly code
  - ▹ Implemented the scanner using **Lex**, the parser using **Yacc**, and constructed the object-oriented Abstract Syntax Tree representation in **C++**, enabling the efficient processing of arithmetic and relational expressions, loops, control flow statements, and function usage
- Custom Shell and Feature Extension of xv6 — Fall'21
  Guide: Prof. Mythili Vutukuru — Operating Systems
  - ▹ Implemented a custom shell with serial, parallel, and background command execution, plus signal handling
  - ▹ Designed and implemented a **priority-based** scheduling algorithm in xv6, improving task execution efficiency
  - ▹ Enhanced xv6 memory management by integrating **lazy page allocation**, improving memory utilization

- Understanding Linux Kernel Internals Through Custom Module Implementation Spring'23
  Guide: Prof. Purushottam Kulkarni Topics in Virtualization and Cloud Computing
  - ▷ Designed kernel modules to explore **kernel internals** with process listing and heap analysis capabilities
  - ▷ Enhanced modules to determine kernel stack pointers, map address spaces, and measure memory allocations
- 3D Visualization and Analysis of Seismic Volumes Spring'23
  Guide: Prof. Prabhu Ramachandran Parallel Scientific Computing and Visualization
  - ▷ Developed a visualization tool using the **Mayavi** and **TraitsUI** Python libraries for interactive geological analysis
  - ▷ Enabled advanced geophysical analysis and multi-dimensional visualization for subsurface investigations
- Justice System and Prison Overflow Spring'23
  Guide: Prof. Om P. Damani System Dynamics: Modeling & Simulation for Development
  - ▷ Conducted a literature survey to identify factors contributing to prison overflow and developed a **system-dynamics model** to simulate population trends, providing insights for reforms to mitigate overcrowding
- Robust Mastermind Player Spring'21
  Guide: Prof. Ashutosh Gupta Logic for Computer Science
  - ▷ Formulated and implemented a logic-based game, Mastermind player, using **SAT**-solving techniques and the Z3 Theorem Prover, which gives accurate performance even with inconsistent and unreliable adversary feedback

Dept. of Computer Science, Virginia Tech

- Two-tier memory management for Compute Express Link (CXL) memory [redacted] – [redacted]
  Guide: Prof. Huaicheng Li Remote
  - ▷ Integrated Data Access MONitor-based memory management patches into Linux and reviewed the source code
  - ▷ Analyzed Redis performance on emulated CXL memory using **YCSB** benchmarks and compared results with vanilla Linux memory management configurations to identify improvements and bottlenecks

TALKS

Compute and Communication trade-offs for scalable Large Language Models (LLMs)
Host: Prof. Purushottam Kulkarni, SynerG Lab, IIT Bombay Jan 2024

AI-Infrastructure Reading Group, Microsoft Research India Lab
Flux: Fast Software-based Communication Overlap On GPUs Through Kernel Fusion Aug 2024
Splitwise: Efficient generative LLM inference using phase splitting Apr 2024

COURSE PROJECTS

Network Simulation Spring'21
Implemented a File Transfer Protocol in **C** and analyzed TCP variants' throughput using Wireshark and NS3

Online Computing and Development Environment (IDE) (Code) Fall'20
Developed a Django-based multi-language online IDE with real-time testing, file storage, and package support

Data Prefetchers and Cache Replacement Interaction (Code) Fall'21
Compared cache policies (LRU, Hawkeye) combined with prefetchers (PACMan, IPCP) across diverse traces

Multi-cycle RISC Processor (Code) Spring'21
Implemented an 8-register, 16-bit multi-cycle processor with sync write and async read operations in VHDL

Real-Time Application Monitor Spring'22
Built a resource monitoring app utilizing Telegraf for data collection and a time-series database for storage

KEY COURSEWORK

Systems and Networking:
Topics in Virtualization and Cloud Computing, Operating Systems, Computer Networks, Parallel Scientific Computing and Visualization, Database and Information Systems, Implementation of Programming Languages, Computer Architecture, Principal of Systems and Data Security, Introduction to GPU Programming (Online)

AI/ML:
Introduction to AI/ML, Foundations of Reinforcement Learning, Automatic Speech Recognition

TECHNICAL SKILLS

Programming: CUDA, Python, C/C++, Java, MATLAB, Bash, SQL, Assembly
Software & Tools: PyTorch, LaTeX, Git, Lex, Yacc, Mayavi, TraitsUI, ChampSim, NS-3
Tools/Frameworks: HTML, CSS, JavaScript, Angular, Django

EXTRA -CURRICULAR ACTIVITIES

National Service Scheme (NSS), IIT Bombay 2019 – 2020
Completed **80+** hours of community service at Social Development under the National Service Scheme
Wrote blogs on sustainable development for Parivartan, an NSS initiative, contributing to its awareness efforts

National Cadet Corps (NCC), Banaras Hindu University 2015 – 2017
Awarded the National Cadet Corps (NCC) 'A' certificate for completing training in the Junior Division Air Wing
Attended the Annual Training Camp-311, NCC, which included rigorous physical training, drills, and sports

Interests: Hindi/Urdu Poetry, Hiking Present